Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

Data and text mining

Supplementary material of NSSRF: global network similarity search with subgraph signatures and its applications

Jiao Zhang, Sam Kwong, Yuheng Jia, Ka-Chun Wong *

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong

*To whom correspondence should be addressed.

Abstract

This supplementary material illustrates the related descriptions, definitions, datasets, and experiment results of NSSRF. In addition, the limitation of NSSRF in terms of network size using five PPI networks to query the random forest regression (RFR) model is discussed. Moreover, the practical memory usage of NSSRF on the four real world datasets are illustrated in this supplementary manuscript.

1 Introduction

Networks are extensively used in bioinformatics, cheminformatics, biomedical, social network analysis, and other application domains (Von Mering *et al.*, 2002; Rual *et al.*, 2005; Szklarczyk *et al.*, 2011; Leskovec and Sosič, 2016; Robinson *et al.*, 2015). The existing database systems face a significant challenge raised by the emergence of massive topological data (Koyutürk *et al.*, 2006). Researchers build network models from various fields, and compare network structure to uncover relationships behind unknown data for different purposes (Hagadone, 1992). Besides the application in bioinformatics, network similarity search (NSS) also plays a key role in social community detection. Analyzing structure within networks provides insights into the functional organization, which in turn contributes knowledge for possible actions, such as the recommendations, and marketing plans for scientific and commercial purposes (Petrakis and Faloutsos, 1997; Willett *et al.*, 1998; Plantié and Crampes, 2013).

Comparative analysis of networks, such as aligning PPI networks across species is network alignment (NA) (Aladağ and Erten, 2013; Dohrmann *et al.*, 2015). However, NA is a NP-complete problem (Döpmann, 2013), which is gaining importance in various domains. Local network alignment (LNA) is measured biologically, such as functional consistency (FC) that indicates the biological relations among vertices using gene ontology (GO) or human phenotype ontology (HPO) terms. NetAligner (Pache *et al.*, 2012), AlignMCL (Mina and Guzzi, 2012), and AlignNemo (Ciriello *et al.*, 2012) are examples of LNA. In biology, the global network alignment (GNA) and similarity search always utilize biological function (Patro and Kingsford, 2012). IsoRankN (Liao *et al.*, 2009) is guided by the intuition that two vertices should be matched if their neighbors are matched, the BLAST scores (Altschul *et al.*, 1990) of sequence similarity between vertices (proteins) are also used in IsoRankN. Graemlin 2.0 constructs GNA based on phylogenetic relationships (Flannick *et al.*, 2006).

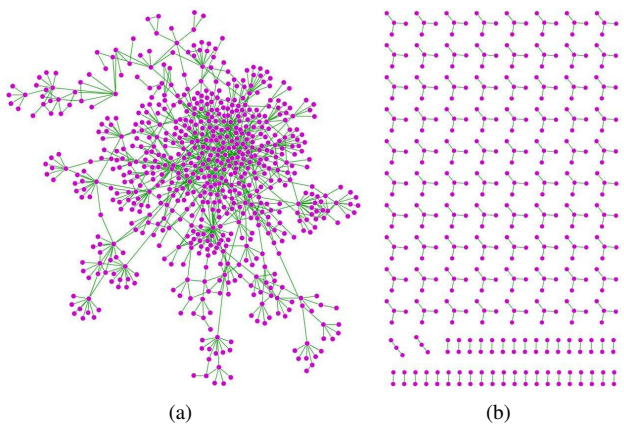
The edge correctness (EC) and largest common connected subgraph (LCCS) are two widely used NA metrics. The NA metrics of two GNA methods are used in our method NSSRF. One GNA method is HubAlign (Hashemifar and Xu, 2014), which adopts a minimum-degree heuristic algorithm to evaluate topology and function importance of protein for PPI networks. Another GNA method is NETAL (Neyshabur *et al.*, 2013), which uses a greedy strategy based on alignment scoring matrix derived from both topological and biological information of input networks.

The usage of network database involved in various areas, and network structure proves to be of utmost significance in NA and NSS. NSS is gaining importance in various areas, TALE (Tian and Patel, 2008), GADDI (Zhang *et al.*, 2009), and NetMatch (Ferro *et al.*, 2007) are focused on subgraph querying. Few works have been done on the global NSS. We propose a global Network Similarity Search method based on Random Forest regression (NSSRF), which has offline model building and similarity query two phases. In the offline model building phase, NSSRF utilizes subgraph signatures and cosine similarity score as features. EC and LCCS of the pairwise NA quality are considered as label to train a model. In the network similarity query phase, each query applies the offline model to predict the EC or LCCS between the query network and the target network. Finally, we can get the similarity score of the query network and the target network in the database by ranking the predicted EC or LCCS, respectively.

2 Methods

2.1 Network and subgraph isomorphism definition

Networks are used to model complex structure with vertices and edges in the chemical compound, metabolic pathway, and PPI networks (Guimera



(a) (b)

Fig. 1. LCCS distinguishes the large contiguous subgraph (a), or small disconnected fragments (b).

Algorithm 1 Offline model training algorithmic pseudo-code of NSSRF.

Input:
$D = \{N_1, N_2, \ldots, N_n\}$: network database
$T = \{T_1, T_2, \ldots, T_j\}$: training networks
Output:
RFR model
for each $N_i \epsilon N$, $T_i \epsilon T$ **do**
normalize subgraph frequency $Sub_k N_i$, $Sub_k T_i$
calculate cosine similarity $S_{Cosk} NT$
get label $L_{EC} NT$, $L_{LCCS} NT$
end for
RFR model training
return RFR model

et al., 2007). Two graphs $N_i = (V_i, E_i)$ and $N_j = (V_j, E_j)$ are isomorphic, denoted by $N_i \cong N_j$, if there is a bijection between the vertex sets of N_i and N_j.

Subgraph isomorphism can be defined as an injection. The problem of determining whether or not a graph $N_i = (V_i, E_i)$ is isomorphic to a subgraph of another graph $N_j = (V_j, E_j)$ is an NP-complete problem, denoted by $N_i \cong S_j \subseteq N_j$. If there is an injection $f: V_i \to V_j$ such that, for each pair of vertices $u_i, v_i \in V_i$, if $(u_i, v_i) \in E_i$ then $(f(u_i), f(v_i)) \in E_j$.

2.2 Network similarity search

Network similarity search (NSS) methods can be divided into two categories. For the sequence based NSS, the similarity between two networks is measured by the number of their common elementary (Willett *et al.*, 1998). For instance, the representation of a network based on sequence alignment is $N = [f_1, f_2, \ldots, f_n]^T$, in which f_i is the number of aligned elements. However, this approach is not accurate due to lacking global topology connectivity (Yan *et al.*, 2005).

For the topology based NSS, the similarity search method is evaluated by structural features (Shapiro and Haralick, 1981). The similarity between two networks is evaluated by the maximal common subgraph. For example, the similarity between two networks N_i and N_j can be defined as $s(N_i, N_j) = \frac{|mcs(N_i, N_j)|}{max(|N_i|, |N_j|)}$, where $|mcs(N_i, N_j)|$ is the value of maximal common subgraph, (Bunke and Shearer, 1998).

Table 1. The network size in terms of vertex and edge in the four real world datasets (AIDS, Bos Taurus, Homo Sapiens I, and Homo Sapiens II).

Dataset	Network Database			Query Networks		
	Network #	Vertex #	Edge #	Network #	Vertex#	Edge #
AIDS	500	3-176	3-182	30	7-24	6-25
Bos Taurus	200	8-360	7-314	30	5-359	5-371
Homo Sapiens I	609	5-855	5-648	30	7-440	6-388
Homo Sapiens II	609	5-347	5-273	30	303-855	291-648

Table 2. The practical memory usage of NSSRF including feature extraction, and offline model training on the four real world datasets in terms of varying subgraph sizes. 234-node indicates using the features of combination of 2-node, 3-node and 4-node subgraph.

dataset	Feature Extraction (MBs)			RFR offine training (MBs)			
	2-node	3-node	4-node	2-node	3-node	4-node	234-node
AIDS	696	696	702	754	770	826	837
Bos Taurus	692	692	699	740	759	844	869
Homo Sapiens I	680	682	686	743	795	1087	1157
Homo Sapiens II	696	698	702	756	812	1116	1180

3 Results

3.1 Datasets

We tested NSSRF on four real world datasets, one of which is molecular dataset. The molecular dataset is the chemical structural from NCI/NIH AIDS Antiviral Screen Data (http://dtp.cancer.gov). The other three datasets are biological pathway datasets, one of which is Bos Taurus pathway dataset, the other two datasets are Homo Sapiens pathway. The difference between Homo Sapiens I and II is the training query networks. The training query networks used in Homo Sapiens I are randomly picked from the dataset, while the training query networks used in Homo Sapiens II are the greatest 30 networks in the dataset. Network size in terms of vertex and edge used in the experiments are listed in Table 1.

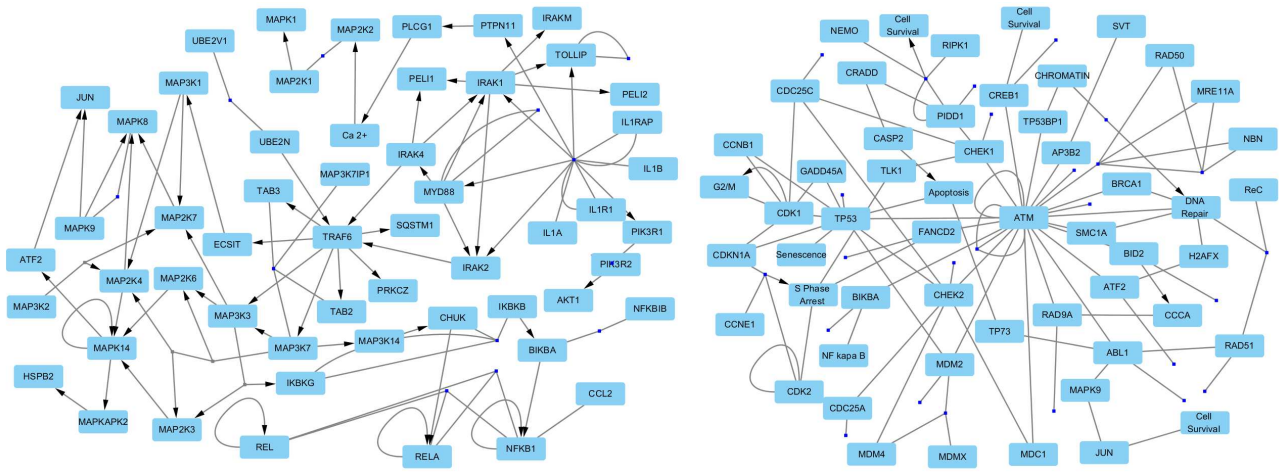
3.2 NA quality measurement: LCCS

The LCCS value of Fig. 1(a) and Fig. 1(b) is 987 and 3, respectively. Larger contiguous subgraph indicates higher similarity, and vice versa. Therefore, for a given query network, if it has the same EC score between the network in Fig. 1(a) and Fig. 1(b); the network in Fig. 1(a) is preferred as similar networks because of its large LCCS value. Formula (1) shows label vectors of EC and LCCS, which are used as labels in the RFR model training phase.

$$\begin{aligned} L_{EC}NT &= [L_{EC}NT_1, L_{EC}NT_2, \ldots, L_{EC}NT_n]^T, \\ L_{LCCS}NT &= [L_{LCCS}NT_1, \ldots, L_{LCCS}NT_n]^T. \end{aligned} \quad (1)$$

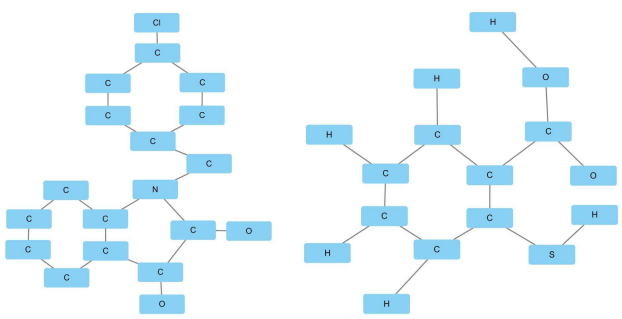
3.3 Comparison of EC and LCCS returned by various methodologies

We have included boxplots to compare the EC and LCCS scores of the various methodologies on the four real world datasets in Fig. 6 and 7, respectively. The black dot on the boxplot is the average EC and LCCS score returned by the corresponding methodology. The statistical significance with a 1% (p<=0.01) significance level of t-test and Wilcoxon's rank-sum test on EC and LCCS are also ascertained on the returned results. Since t-test has the same result with Wilcoxon's rank-sum test, only the Wilcoxon's rank-sum test is denotated on the figures. NSSRF marked with * and ˆ denotes that the performance of NSSRF is significantly better than C-tree and SIGMA.



(a) (b)

Fig. 2. Case studies of NSSRF using LCCS of NETAL as similarity metric on the Bos Taurus dataset from WikiPathways website (Kelder et al., 2012). (a) is IL-1 signaling pathway WP3271; (b) is ATM signaling pathway WP3221. (a) and (b) are drawn by network visualization tool Cytoscape (Shannon et al., 2003).



(a) (b)

Fig. 3. Case studies of NSSRF using EC of NETAL as similarity metric on AIDS Antiviral Screen Data. (a) and (b) are the query chemical compound #502, and the chemical compound #100 in the AIDS database, respectively. (a) and (b) are drawn by network visualization tool Cytoscape (Shannon et al., 2003)

3.4 Time cost on WikiPathways datasets

Time cost on three WikiPathways datasets are shown in Fig. 8, which is the average time cost of 30 query networks under 10-fold cross-validation. For the WikiPathways datasets, the offline model training time of NSSRF is comparable with C-tree. However, the query phase of NSSRF is faster than C-tree in all the cases.

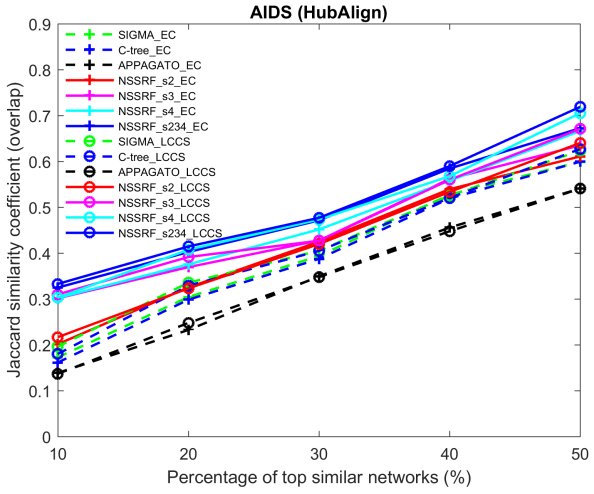
3.5 Limitations of NSSRF in terms of network size

The network size is expressed in terms of the number of nodes and edges. In this study, the network sizes on the four datasets are listed in Table 1. The numbers of both the nodes and edges are below 1000. For a network with less than 1000 nodes and edges, it needs less than a second to extract all the 2-node, 3-node and 4-node subgraphs using Mfinder. Since we have reduced the network to subgraphs in the feature vectors, and totally there are 2 different types of directed 2-node subgraphs, 13 3-node subgraphs and 199 4-node subgraphs. The training time of 10-fold cross-validation takes less than 5 seconds, the query time takes less than 0.002 seconds on the AIDS, Bos Taurus, Homo Sapiens I and Homo Sapiens II dataset. Therefore, NSSRF does not have any practical restrictions in terms of network size on both the regression model building and query stage on the AIDS compound networks and WikiPathways networks in this study.

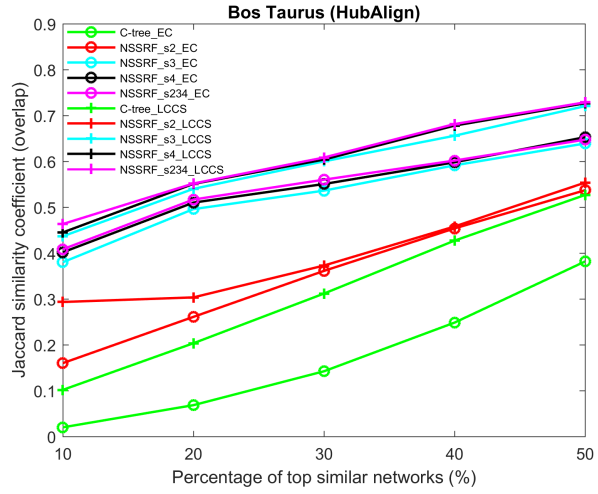
Table 3. PPI networks querying Homo Sapiens I dataset. There are 609 networks in the database, and 30 networks are used as query network in the training stage. The training time of 2-node, 3-node and 4-node is 0.01, 0.6 and 1.7 seconds, respectively.

PPI	Vertex #	Edge #	Feature Extraction Time (seconds)			Query Time (seconds)		
			2-node	3-node	4-node	2-node	3-node	4-node
Mouse	288	242	1×10^{-4}	1×10^{-4}	0.4	2×10^{-4}	9×10^{-4}	1×10^{-3}
Worm	2795	4495	2×10^{-4}	8×10^{-4}	21	2×10^{-4}	1.1×10^{-3}	1.2×10^{-3}
Fly	7510	25635	1×10^{-3}	7	376	5×10^{-4}	1.2×10^{-3}	1.2×10^{-3}
Yeast	5495	31261	1	10	1026	5×10^{-4}	1.2×10^{-3}	1.4×10^{-3}
Human	9476	34327	1	14	1192	6×10^{-4}	1.2×10^{-3}	1.5×10^{-3}

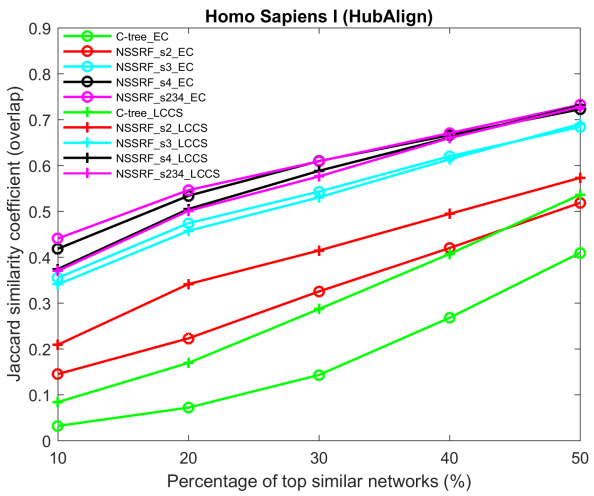
However, for large PPI networks, NSSRF needs a relatively long time in the feature extraction stage due to the NP-complete problem of the subgraph isomorphism. In order to compare the runtime of NSSRF on varying size networks, five PPI networks (mouse, worm, fly, yeast, and human) which have been tested in IsoRankN (Liao *et al.*, 2009) are tried using NSSRF on the Homo Sapiens I dataset. The number of nodes/edges of the five PPI query networks and corresponding feature extraction and query time are tabulated in Table 3. There are 609 networks in the Homo Sapiens I dataset. LCCS generated by NETAL are used as labels. The training time for using features of 2-node, 3-node and 4-node subgraphs are 0.01, 0.6 and 1.7 seconds, respectively. The network sizes of the five PPI networks do not have any major effect on the query time cost. However, the feature extraction time of 4-node subgraphs vary with increasing number of edges; for instance, the PPI network of Human with 9476 nodes and 34327 edges needs approximately 19.86 minutes (i.e. 1192 seconds) to extract all 4-node subgraphs. Therefore, the subgraph extraction is the limiting step in NSSRF due to the NP-completeness of subgraph isomorphism to some extent.



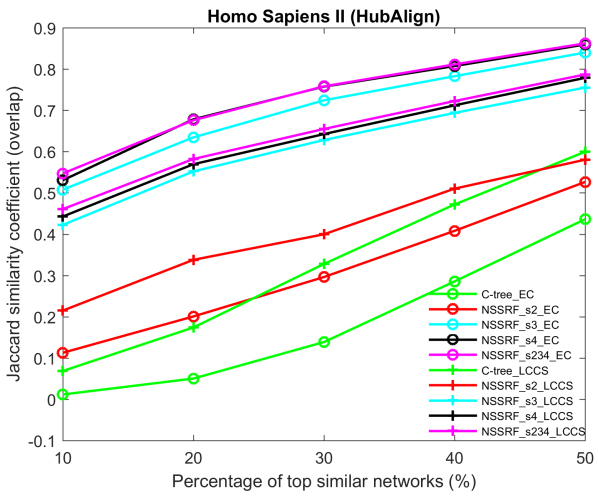
(a)



(b)

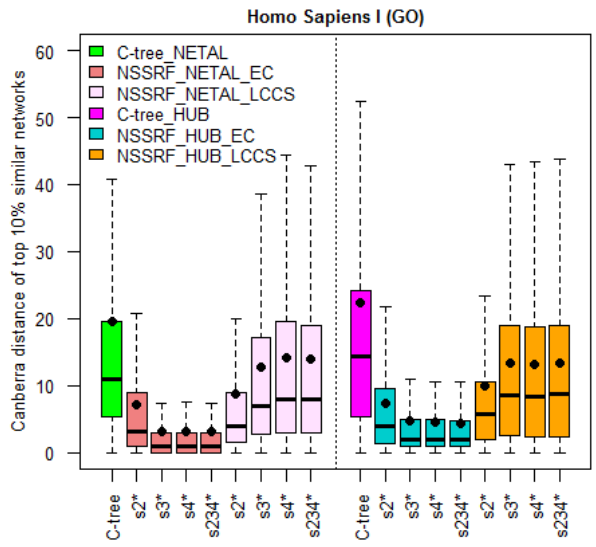


(c)

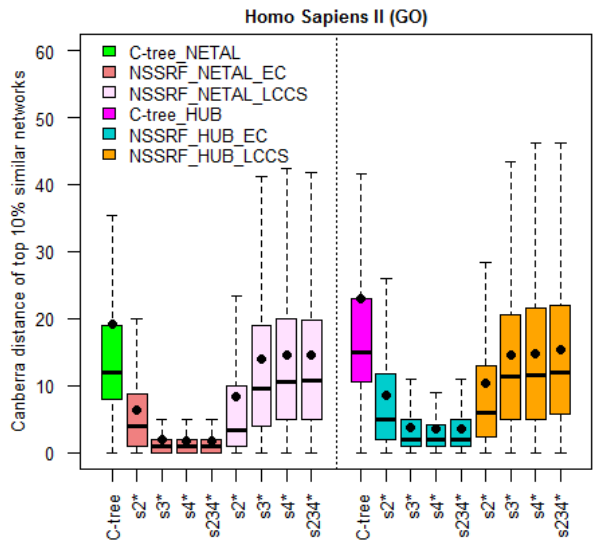


(d)

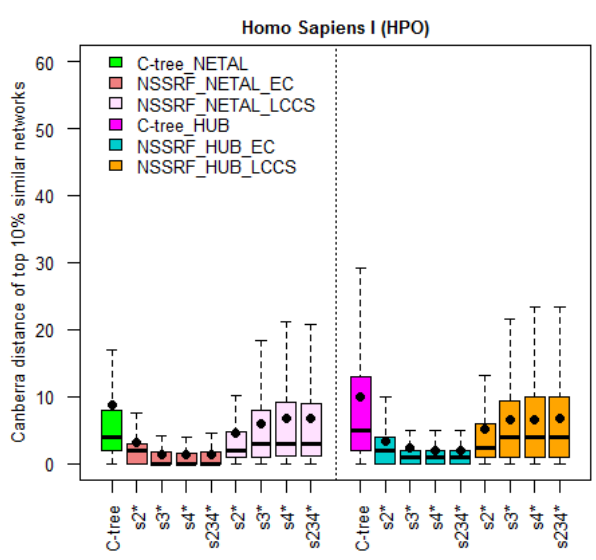
Fig. 4. The overlap performance comparison of Jaccard similarity coefficient on AIDS, Bos Taurus, Homo Sapiens I, and Homo Sapiens II datasets under 10-fold cross-validation. EC and LCCS generated by HubAlign is used. _EC and _LCCS indicate evaluating the performance of NSSRF on EC and LCCS, respectively. NSSRF_s2_EC indicate evaluating NSSRF with 2-node subgraph on EC, which is the same as the other subgraph sizes used in NSSRF. Note: Only C-tree and NSSRF can be run on the WikiPathways datasets.



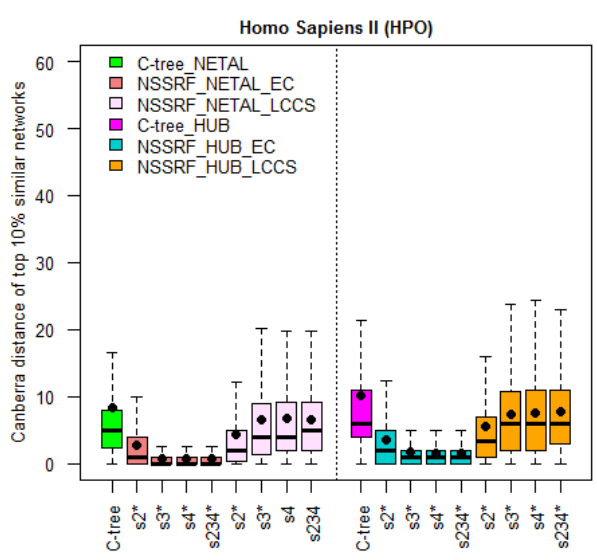
(a)



(b)

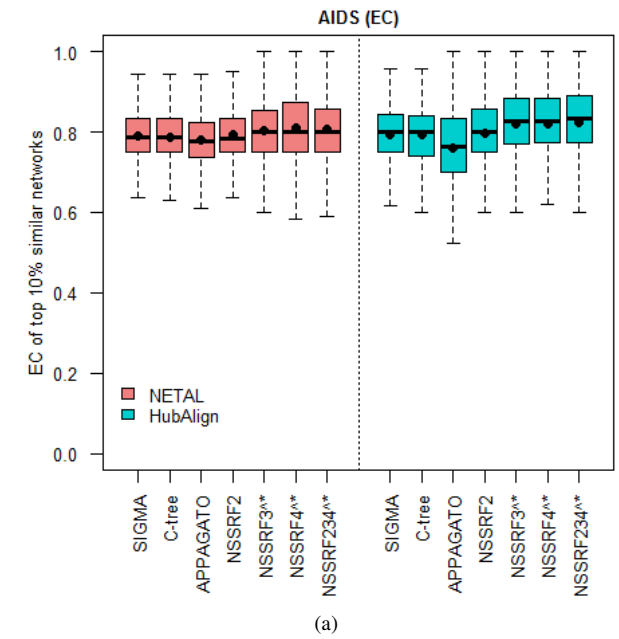


(c)

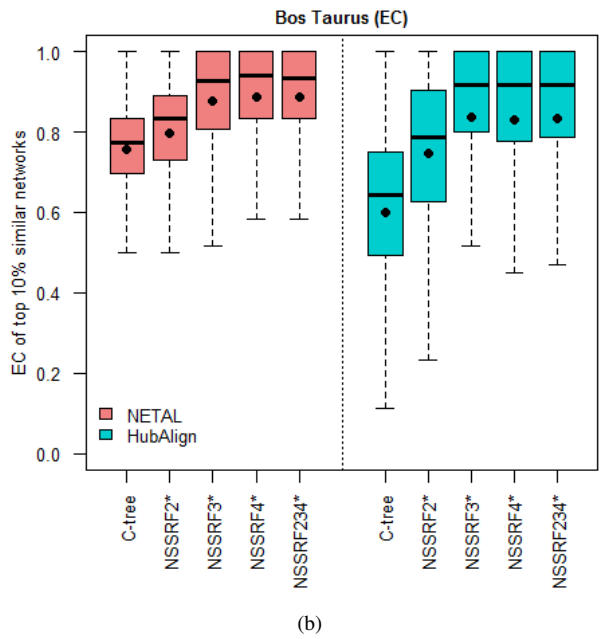


(d)

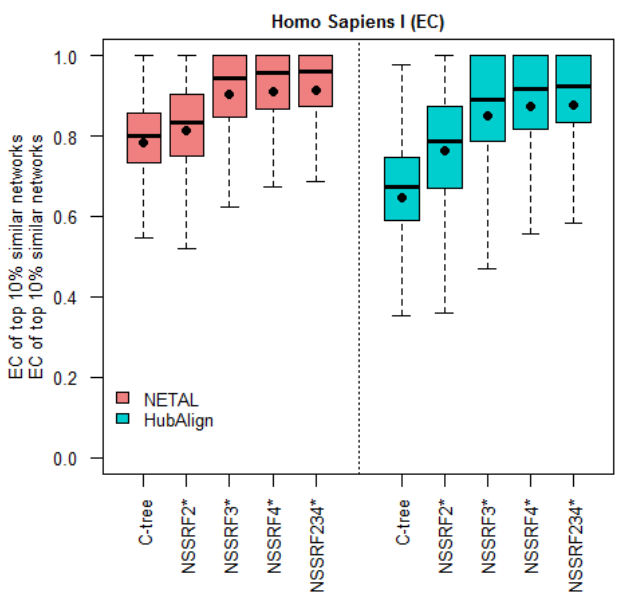
Fig. 5. The biological function distance comparison using Canberra distance to evaluate the methodology from genotype coherence to phenotype aspects on Homo Sapiens I, and Homo Sapiens II dataset. Wilcoxon's rank-sum test with a 1% significance level is conducted. The black dot on the boxplot is the average Canberra distance in terms of GO or HPO terms returned by the corresponding methodology. NSSRF marked with asterisk denotes that the performance of NSSRF is significantly better than C-tree. C-tree_HUB and C-tree_NETAL indicate evaluating the distance of returned networks from C-tree by HubAlign and NETAL, respectively. s2* indicate NSSRF using 2-node subgraph, which is similar to other node sizes. Note: SIGMA and APPAGATO cannot be run on these datasets.



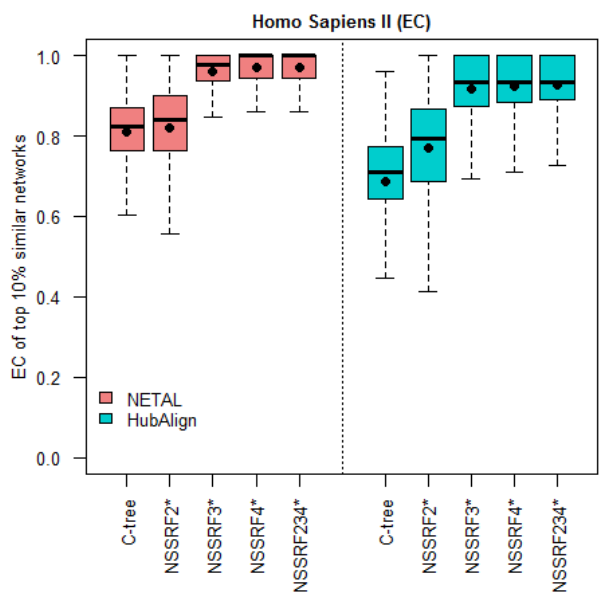
(a)



(b)



(c)



(d)

Fig. 6. The EC returned by various methodologies on AIDS, Bos Taurus, Homo Sapiens I, and Homo Sapiens II datasets under 10-fold cross-validation. Pairwise GNA method NETAL and HubAlign are used to get the alignment quality metrics. NSSRF2 indicate evaluating NSSRF with 2-node subgraph, which is the same as the other sizes used in NSSRF. The black dot on the boxplot is the average EC scores returned by the corresponding methodology. Wilcoxon's rank-sum test with a 1% significance level is conducted. NSSRF marked with * and ^ denote that the performance of NSSRF is significantly better than C-tree and SIGMA, respectively. We did not add any notations on the figure for the case that NSSRF performs significantly better than APPAGATO, because NSSRF outperforms APPAGATO significantly in all the cases. Note: Only C-tree and NSSRF can be run on the WikiPathways datasets.

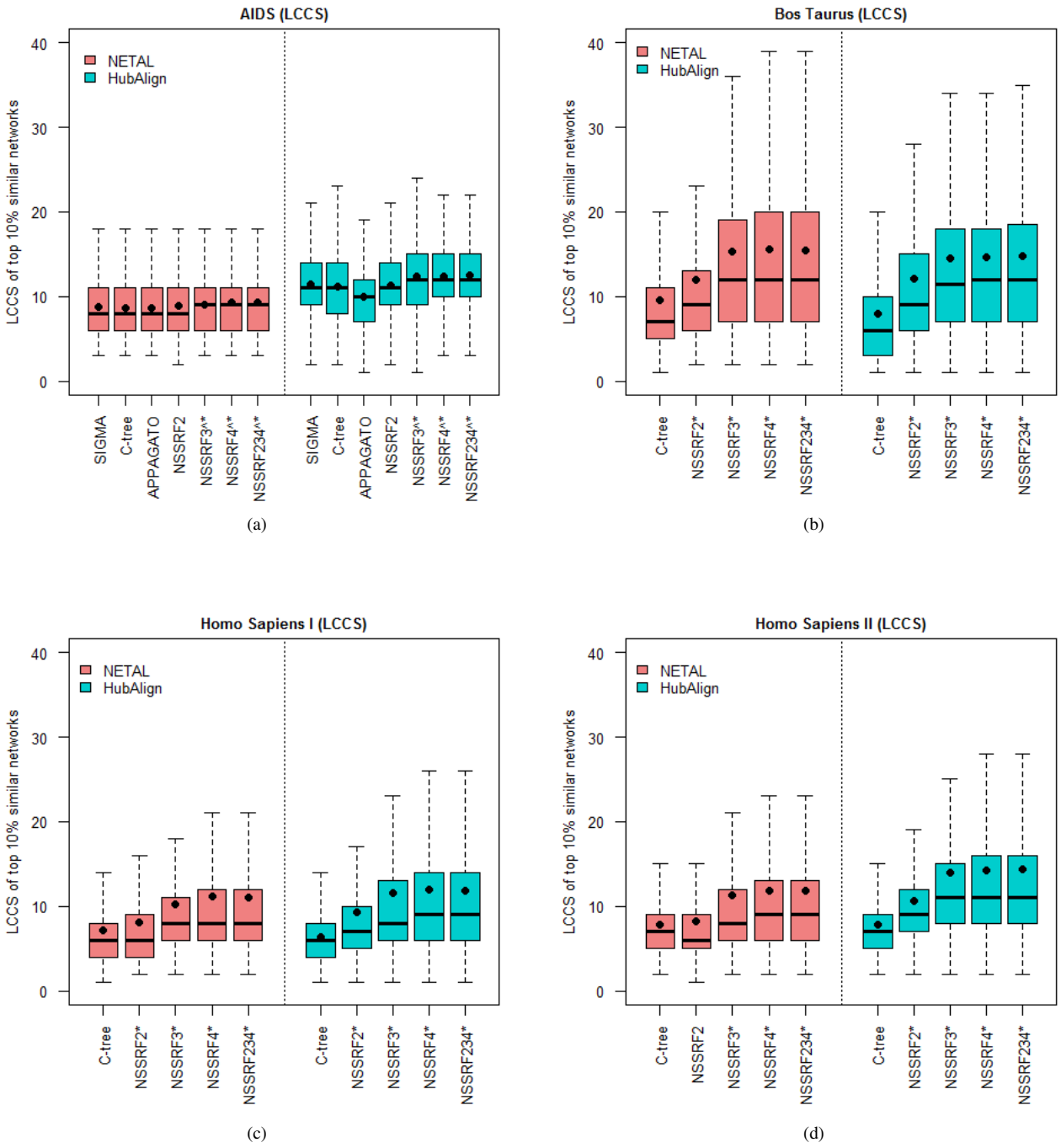
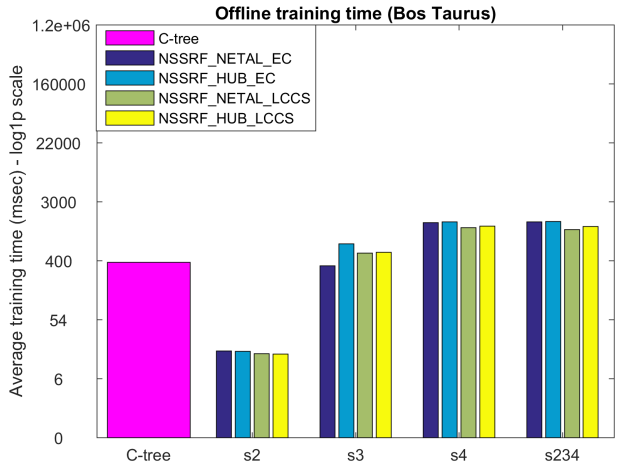
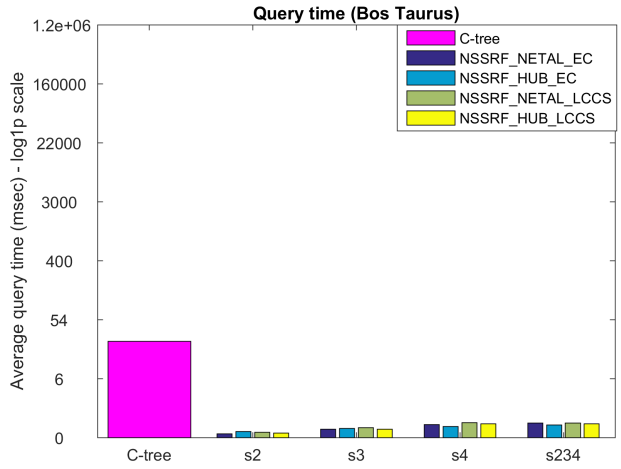


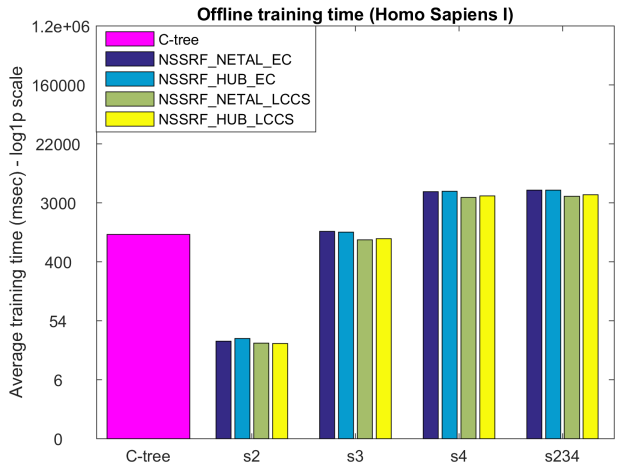
Fig. 7. The LCCS returned by various methodologies on AIDS, Bos Taurus, Homo Sapiens I, and Homo Sapiens II datasets under 10-fold cross-validation. Pairwise GNA method NETAL and HubAlign are used to get the alignment quality metrics. NSSRF2 indicate evaluating NSSRF with 2-node subgraph, which is the same as the other sizes used in NSSRF. The black dot on the boxplot is the average EC scores returned by the corresponding methodology. Wilcoxon's rank-sum test with a 1% significance level is conducted. NSSRF marked with * and ^ denote that the performance of NSSRF is significantly better than C-tree and SIGMA, respectively. We did not add any notations on the figure for the case that NSSRF performs significantly better than APPAGATO, because NSSRF outperforms APPAGATO significantly in all the cases except for NSSRFs2 using LCCS label generated by NETAL. Note: Only C-tree and NSSRF can be run on the WikiPathways datasets.



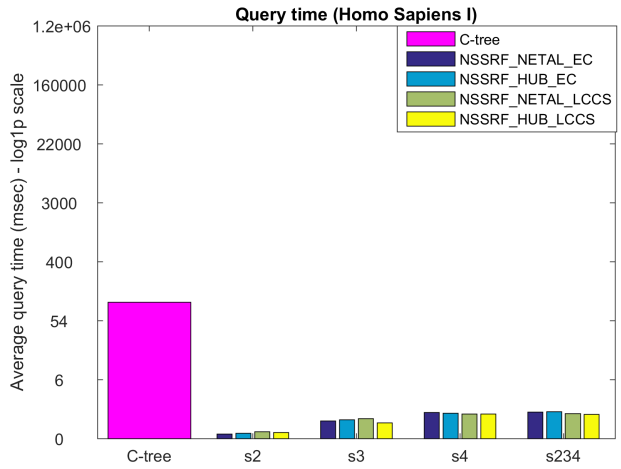
(a)



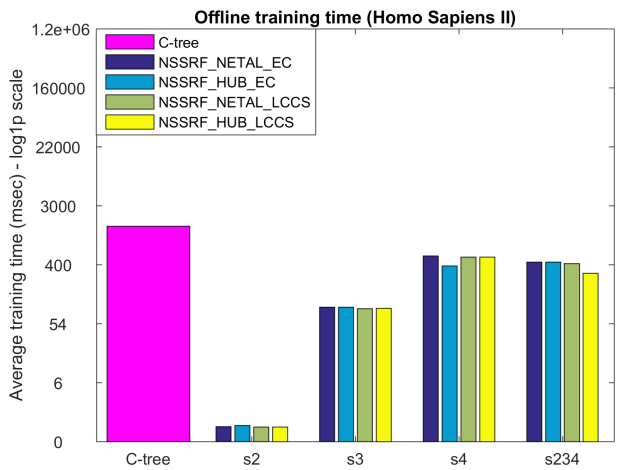
(b)



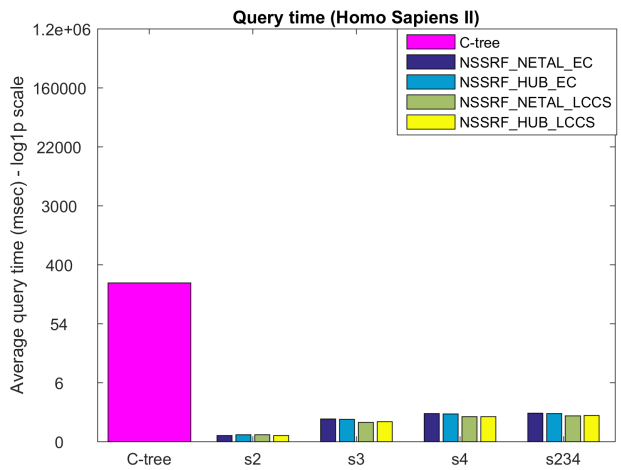
(c)



(d)



(e)



(f)

Fig. 8. The offline training and similarity query time comparison of C-tree and NSSRF on WikiPathways datasets (Bos Taurus, Homo Sapiens I, and Homo Sapiens II.) under 10-fold cross-validation. EC and LCCS generated by NETAL and HubAlign are used. _EC and _LCCS indicate evaluating the performance of NSSRF on EC and LCCS, respectively. s2 indicate evaluating NSSRF with 2-node subgraph, which is the same as the other subgraph sizes used in NSSRF. Notes: SIGMA and APPAGATO cannot be run on these datasets. $log1p$ computes $ln(1+x)$ which is accurately for small values of x.

References

Aladağ, A. E. and Erten, C. (2013). Spinal: scalable protein interaction network alignment. *Bioinformatics*, **29**(7), 917–924.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403–410.

Bunke, H. and Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern recognition letters*, **19**(3), 255–259.

Ciriello, G., Mina, M., Guzzi, P. H., Cannataro, M., and Guerra, C. (2012). Alignnemo: a local network alignment method to integrate homology and topology. *PloS one*, **7**(6), e38107.

Dohrmann, J., Puchin, J., and Singh, R. (2015). Global multiple protein-protein interaction network alignment by combining pairwise network alignments. *BMC bioinformatics*, **16**(Suppl 13), S11.

Döpmann, C. (2013). Survey on the graph alignment problem and a benchmark of suitable algorithms. *Institut für Informatik*.

Ferro, A., Giugno, R., Pigola, G., Pulvirenti, A., Skripin, D., Bader, G., and Shasha, D. (2007). Netmatch: a cytoscape plugin for searching biological networks. *Bioinformatics*, **23**(7), 910–912.

Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H., and Batzoglou, S. (2006). Graemlin: general and robust alignment of multiple large interaction networks. *Genome research*, **16**(9), 1169–1181.

Guimera, R., Sales-Pardo, M., and Amaral, L. A. (2007). Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics*, **3**(1), 63–69.

Hagadone, T. R. (1992). Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *Journal of chemical information and computer sciences*, **32**(5), 515–521.

Hashemifar, S. and Xu, J. (2014). Hubalign: an accurate and efficient method for global alignment of protein–protein interaction networks. *Bioinformatics*, **30**(17), i438–i444.

Kelder, T. *et al.* (2012). Wikipathways: building research communities on biological pathways. *Nucleic acids research*, **40**(D1), D1301–D1307.

Koyutürk, M., Kim, Y., Topkara, U., Subramaniam, S., Szpankowski, W., and Grama, A. (2006). Pairwise alignment of protein interaction networks. *Journal of Computational Biology*, **13**(2), 182–199.

Leskovec, J. and Sosič, R. (2016). Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **8**(1), 1.

Liao, C.-S., Lu, K., Baym, M., Singh, R., and Berger, B. (2009). Isorankn: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, **25**(12), i253–i258.

Mina, M. and Guzzi, P. H. (2012). Alignmcl: Comparative analysis of protein interaction networks through markov clustering. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, pages 174–181. IEEE.

Neyshabur, B. *et al.* (2013). Netal: a new graph-based method for global alignment of protein–protein interaction networks. *Bioinformatics*, **29**(13), 1654–1662.

Pache, R. A., Céol, A., and Aloy, P. (2012). NetalignerâĽ"a network alignment server to compare complexes, pathways and whole interactomes. *Nucleic acids research*, **40**(W1), W157–W161.

Patro, R. and Kingsford, C. (2012). Global network alignment using multiscale spectral signatures. *Bioinformatics*, **28**(23), 3105–3114.

Petrakis, E. G. M. and Faloutsos, A. (1997). Similarity searching in medical image databases. *IEEE Transactions on Knowledge and Data Engineering*, **9**(3), 435–447.

Plantié, M. and Crampes, M. (2013). Survey on social community detection. In *Social media retrieval*, pages 65–85. Springer.

Robinson, I., Webber, J., and Eifrem, E. (2015). *Graph Databases: New Opportunities for Connected Data*. " O'Reilly Media, Inc.".

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., *et al.* (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**(7062), 1173–1178.

Shannon, P. *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**(11), 2498–2504.

Shapiro, L. G. and Haralick, R. M. (1981). Structural descriptions and inexact matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5), 504–519.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., *et al.* (2011). The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, **39**(suppl 1), D561–D568.

Tian, Y. and Patel, J. M. (2008). Tale: A tool for approximate large graph matching. In *2008 IEEE 24th International Conference on Data Engineering*, pages 963–972. IEEE.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**(6887), 399–403.

Willett, P. *et al.* (1998). Chemical similarity searching. *Journal of chemical information and computer sciences*, **38**(6), 983–996.

Yan, X., Yu, P. S., and Han, J. (2005). Substructure similarity search in graph databases. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 766–777. ACM.

Zhang, S., Li, S., and Yang, J. (2009). Gaddi: distance index based subgraph matching in biological networks. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 192–203. ACM.